Vanda Broughton  (University College London)

Finding Bliss on the Web
Some problems of representing faceted terminologies in digital environments

Abstract
The Bliss Bibliographic Classification is the only example of a fully faceted general classification scheme in the Western world. Although it is the object of much interest as a model for other tools it suffers from the lack of a web presence, and remediying this is an immediate objective for its editors. Understanding how this might be done presents some challenges, as the scheme is semantically very rich and complex in the range and nature of the relationships it contains. The automatic management of these is already in place using local software, but exporting this to a common data format needs careful thought and planning. Various encoding schemes, both for traditional classifications, and for digital materials, represent variously: the concepts; their functional roles; and the relationships between them. Integrating these aspects in a coherent and interchangeable manner appears to be achievable, but the most appropriate format is as yet unclear.

1: Introduction
Facet analysis remains virtually alone as a rational intellectual method for the construction of subject terminologies. Developed for the linear arrangement of items in libraries, the rigour of its analytical approach, and the clear delineation of relationships between concepts in a subject domain make it an excellent contender for the management of subject content in a digital environment. The faceted classification, of which the second edition of the Bliss bibliographic classification (BC2) (Mills & Broughton 1977) is the most well developed example, fulfils all the necessary requirements of a well structured indexing tool: the use of standard fundamental categories to accommodate concepts; a well-expressed methodology for ordering and organization within categories, clear specification of relationships between terms, both intra- and inter-category; and a logical and predictable system syntax.

2: Coding in electronic BC2
BC2 is maintained in an electronic format, in which all the structural components of the terminology can be expressed in a machine readable manner, although this format is currently not suitable for a public interface. BC2 uses an “in-house” system of mark-up, developed some years ago for creating enhanced classification data, using a basic text editor. It is in many ways comparable with XML in that the principle purpose is to identify the structural and semantic nature of concepts in the terminology, rather than the formatting elements associated with HTML; although, in the first instance, the mark-up was created to generate the physical display of the classification in the published schedules, its purpose was to control the layout of the schedule in terms of the visual representation of facets and hierarchies.

Coding of concepts in this way, in classifications and in other controlled indexing languages, has a well established history, with some examples going back to the mid-20th century (Grolier 1962). Two main types of coding can be identified: coding of concepts with role indicators (sometimes specifically identifiable as facet indicators, as is the case with the UDC and with the Colon classification) so that the functional status of the concept can be recognized; or, coding of the relationships between concepts. The
latter is best represented in the UK tradition by the work of Jason Farradane in his
system of relational indexing (Farradane 1950), where the concepts, or isolates, were
neutral in status, and linked by relationship codes.

To some extent the use of role or facet indicators represents a form of faceted
classification practice alternative to that exemplified by the United Kingdom
Classification Research Group, where the ordering of facets in the linear sequence, and
the combination of concepts between facets, is controlled by the notion of a standard
citation order. BC2 in its original print format contained no coding (other than the
notation), and the implicit system syntax, applied intellectually, generates the
complexity of structure in a logical and consistent manner. It provides no basis for
machine recognition or manipulation, since there is simply no data present which might
easily be translated to a machine readable format.

In the electronic version of BC2, the encoding is of a more narrowly structural
nature, but nevertheless supports the identification of a range of relationships (and the
inference of their reciprocals) by the system software. Characteristics of concepts
(classes) that are encoded in various ways include:

- position in the linear sequence of concepts
- hierarchical position
- “non-concept” status, which is used for elements such as node labels and other
  “signposts” which do not themselves constitute classes as such
- editorial information, for example, whether terms are to be included in the
  alphabetical index to the classification.

On the basis of this encoding, the software is able to infer additional information,
notably about the relationships between concepts. Since the program knows the relative
position of concepts both in the linear order and in the hierarchy, and it can detect when
“non-classes” interrupt the sequence, it is possible automatically to determine the
semantic/paradigmatic relations, i.e. relationships of super- and sub-, and co-ordination
(or in thesaurus or database terms, broader, narrower and some associative terms, and
parent-child and sibling relationships).

3: Thesaurus generation from the code
BC2 has long shown potential for conversion to the thesaurus format, and this has
begun to be realised (Aitchison 2004). Recent work on developing and extending the
software has supported the semi-automatic generation of a compatible thesaurus from
the same input files that are used for the classification display, and the alphabetical
index (Broughton 2008a). There are some significant problems with vocabulary control,
since the original classificatory structure had no need to consider any potential use of
class headings or captions as indexing terms or lexical labels; concepts in the
classification scheme are represented by the notation, and there is no need to look for
economy or elegance in the way in which they are described. Current work on the
development of editorial guidelines (Aitchison 2009) should alleviate this problem in
the future. Otherwise, some degree of manual editing of the program output is required,
and the software cannot as yet cope with the majority of associative terms. This is
because they represent interaction of concepts across categories, rather than within
categories, and the present mark-up does not provide information about categorical
status i.e. it does not use role or facet indicators to show whether concepts are entities,
parts, processes, agents, etc. In the example of code shown in Figure 1, in addition to the encoding for features such as hierarchy, class validity and index inclusion, related terms are added manually to the schedule. It does however, seem a logical step to incorporate this further characteristic into the mark-up language.

Figure 1: BC2 schedule for Class J “Education” in the encoded format

To a limited extent therefore, the software acts as a bridge between the encoding of concepts and the automatic generation of relationships which do not in themselves need to be specified or encoded. The software is essentially an inference engine for the BC2 ontology.

4: Encoding of facets in the FATKS project

Work on the FATKS project involved the creation of a relational database to hold faceted classification data (Broughton & Slavić 2007). Many of the features of the BC2 mark-up tool are replicated in the database, but it has much greater functionality than the mark-up and can, for instance, deal with the system syntax, and allow for automatic combination of terms and generation of classmarks (which it is difficult to emulate with the BC2 programs). This is a direct result of the encoding of category status, which, through the medium of citation order, allows a default ordering and combination of concepts (FATKS n.d.).
In this example, the J and (K) notation encodes the status of operation and agent, and the numerical code indicates entity; the three are combined in default order to create a compound notation for the compound concept. Were this to be done in the BC2 mark-up, it would facilitate to a degree the identification of a wider range of associative terms, particularly where these are represented through examples of combination in the BC2 structure.

5: The role of encoding
The adequate representation of relationships underpins a number of important functions of a subject tool. Although concepts can be used in a post-coordinate manner to tag documents for retrieval, concepts are more usefully represented when they are also part of a systematic structure which facilitates browsing and navigation (either to permit overview of the subject field, or, more narrowly, to allow query formulation and modification in the search function). A general purpose subject tool needs to embrace these features of search and retrieval as well as the more limited indexing role. Because of the range of relationships in the faceted terminology, and the precise nature of those relationships, it provides a more sophisticated modelling of a subject domain, and a correspondingly more effective tool for browsing and navigation as well as query formulation and modification. Careful thought needs to be given, therefore, to the way in which the relationships are formally expressed, or whether they should be derived from other data elements.

Since BC2 is the only general classification currently without a Web-based format (although parts of it are available online as Word documents), it was decided to examine the ways in which it might be represented in a web environment. Because the BC2 mark-up is not a suitable format, being unintelligible to browsers and to (uninformed) humans alike, the classification data needs to be represented in some other, more interoperable, way. Current work on BC2 is examining the way in which encoding can be used not only to create a version of the classification for dissemination on the Web, but also to represent the potential complexity of its semantic structure.

BC2 is particularly rich in inter-facet relationships (for example, entity-process, entity-operation, agent-operation, etc.), which the new British standard for structured vocabularies examines as a way of improving precision in the naming of associated relationships (BSI 2005a-b). However, relatively few of these potential connexions are currently expressible in the standards for subject representation, either those for bibliographic use, or for web representation.
Examination of various web representation languages show them to be at least partially deficient in terms of the structural elements of BC2. Most are able only to express relatively simple hierarchical relationships, and cannot reflect the richness of the BC2 structure with its great variety of paradigmatic and syntagmatic interconnections. Even systems such as Extensible faceted markup language (XFML) can handle only a relatively limited range of relationships, principally those which exist within categories, and indeed, it is not untypical of systems that regard a faceted structure as a series of comparable hierarchies representing attributes of entities, rather than the whole range of canonical fundamental categories as found in the LIS model of faceted classifications. The introduction to XFML states specifically that it is intended as “a simple standard to exchange faceted, hierarchical metadata” (Van Dijck 2003), suggesting a relatively unsophisticated metadata tool.

6: BC2 and SKOS (Simple Knowledge Organization System)

Another candidate systems such as SKOS (Simple Knowledge Organization System) was regarded initially as unable to handle more than the simplest structures in a faceted terminology, and the ways in which the range of relationships could be extended offered a substantial challenge. However, the influence of the controlled vocabulary community is very evident in more recent versions of the SKOS documentation, where compatibility with good practice as demonstrated in BS8723 is clearly an objective (Isaac & Summers 2009).

SKOS now addresses a number of the issues traditionally tackled by the thesaurus compiler. The phenomenon of synonymy is managed through the use of lexical labels, which are variously expressed as:

\[
\text{skos:prefLabel} \quad \text{skos:altLabel} \quad \text{skos:hiddenLabel}
\]

to accommodate preferred terms, non-preferred terms, and to allow the retrieval of commonly misspelled items.

Broader and narrower terms are managed through the use of \text{skos:broader} and \text{skos:narrower}, although the fact that many so-called hierarchies are not accurate in their delineation of these relationships is acknowledged by the need separately to specify transitivity in the broader/narrower chain. The use of \text{skos:broaderTransitive} and \text{skos:narrowerTransitive} allows for navigation and query expansion in a “clean” hierarchy. Associative relationships employ the property \text{skos:related}, which is by definition intransitive. There is also provision for representing the use of node labels, which mirror the organization of concepts in arrays by the application of principles of division in the faceted classification. This is handled by the use of the \text{skos:Collection} property in conjunction with \text{skos:member}. Where concepts in array are in a particular order, this too can be represented.

Some degree of automatic management is achieved by the use of an inference engine which can create reciprocals of both \text{skos:broader} and \text{skos:narrower}, and \text{skos:related}. These functions in SKOS provide for most of what can be managed by the BC2 software, although curiously SKOS operates as the inverse of the BC2 programs. In BC2 the structure is encoded and the relationships inferred from the coding: there is a classification→thesaurus conversion. In SKOS, the relationships are
represented, and hierarchies can be generated from that data: there is a thesaurus →
classification conversion. While there is clearly a conceptual compatibility here, it is
not clear whether any automated data conversion could be achieved.

There is however, no means of combining concepts in SKOS, either pre- or post-
coordinately, except where compound terms are routinely managed. Nor is there any
current means of representing inter-facet relationships (as opposed to the semantic intra-
facet relationships) and although there is scope for extending SKOS to meet the
conditions of a particular KOS, it is not very clear at present how much work this would
involve.

7: BC2 and TEI (Text encoding initiative)
Other appropriate work appears to exist in the field of document mark-up where
existing techniques for representing texts, notably those associated with the Text
encoding initiative (TEI), appear to have much in common with facet analysis as a
methodological approach (Broughton 2008b). The functional nature of the tags used in
Figure 3, such as <anthology>, <gram:nom>, <gram:adj>, and <gram:aux>, show
a closer conceptual correspondence to the facet markup codes of FATKS and the BC2
structural codes, than do the properties of SKOS. These techniques, which support
internal analysis and content representation, show potential for transferring to the
structuring of metadata.

Figure 3: Encoding a text using a TEI schema (TEI 2009)

Text encoding tools have been developed for a variety of document types, although
nothing comparable to a large metadata structure like BC2 has yet been attempted.
However, current co-operative work with scholars in the area of humanities computing
suggests that, in combination, facet analytical and text encoding methods may offer a
solution to improving the usability of metadata tools and providing more subtle and
sophisticated means of subject representation.

References
Aitchison J., 2004, Thesauri from BC2: problems and possibilities revealed in an experimental
thesaurus derived from the Bliss Music schedule, Bliss classification bulletin, 46, p. 20-26.


FATKS: Facet analytical theory in knowledge structures. Project details and documentation <www.ucl.ac.uk/fatks/>.

Grolier E. de, 1962, A study of general categories applicable to classification and coding in documentation, Unesco, Paris.


Web documents have been accessed 1 September 2009.